ВЫДЕЛЕНИЕ СМЫСЛОВЫХ ПОНЯТИЙ В МЕДИЦИНСКИХ ДИАГНОЗАХ ПРИ ПОМОЩИ МАШИННОГО ОБУЧЕНИЯ

Д.И. Коваль, И.В. Сушков, А.Б. Тепляков
(г.Томск, Томский Политехнический Университет)
e-mail: deniskoval12@gmail.com, ivs47@tpu.ru, abt4@tpu.ru

ALLOCATION OF SEMANTIC CONCEPTS IN MEDICAL DIAGNOSES BY MEANS OF MACHINE LEARNING

D.I. Koval, I.V. Sushkov, A.B.Teplyakov
(Tomsk, Tomsk Polytechnik University)

Abstract. Biomedical named entity recognition (Bio-NER) is a fundamental task in handling biomedical text terms, such as RNA, protein, cell type, cell line, and DNA. Bio-NER is one of the most elementary and core tasks in biomedical knowledge discovery from texts. The system described here is developed by using the BioNLP/NLPBA 2004 shared task.

Keywords: machine learning methods, natural language analysis, morphology, syntax, semantics, ontology.

“Изучив множество публикаций и исследований в области применения методов машинного обучения на основе нейронных сетей, мы выделили несколько наиболее перспективных, на наш взгляд, направлений в создании и развитии систем искусственного интеллекта для здравоохранения”:

1. **“Автоматизированные системы диагностики**, например, системы для автоматического анализа рентгенологических или МРТ-снимков на предмет выявления патологии, микроскопический анализ биологического материала, автоматическая расшифровка ЭКГ, электроэнцефалограмм и т.д.
2. **Системы распознавания неструктурированных медицинских записей и понимания естественного языка** могут оказать существенную помощь, как врачу, так и пациенту. Начиная от уже обычной расшифровки речи и превращении ее в текст в качестве более продвинутого интерфейса общения с медицинскими информационными системами (МИС), обращения в Call-центр или голосового помощника.
3. **Системы анализа и предсказания событий** также являются вполне решаемыми уже сейчас задачами для ИИ, которые могут дать существенный эффект. Например, оперативный анализ изменений заболеваемости позволяет быстро предсказать изменение обращаемости пациентов в медицинские организации или потребность в лекарственных препаратах.
4. **Системы автоматической классификации и сверки информации** помогают связать информацию о пациенте, находящейся в различных формах в различных информационных системах. Например, построить интегральную электронную медицинскую карту из отдельных эпизодов, описанных с разной детальностью и без четкого или противоречивого структурирования информации.
5. **Автоматические чат-боты для поддержки пациентов** могут оказать существенную помощь в повышении приверженности пациентов здоровому образу жизни и назначенному лечению” [1].

“Исследования в области разработки программного обеспечения для задач обработки естественного языка (Natural Language Processing –NLP, Language Engineering – LE) активно развиваются в различных исследовательских парадигмах. Устойчивые тенденции последнего десятилетия в области LE связаны с широкомасштабными исследованиями в области разработки и применения статистических методов и методов машинного обучения (Machine Learning – ML). Характерными чертами таких исследований являются:

- Использование эмпирических методов с точными критериями оценок

- Расширение сферы применения статистических методов
- Использование больших ресурсов данных (текстовые базы данных, онтологии, тезаурсы, корпуса текстов)
- Применение NLP-технологий в реальных областях”

Можно выделять ряд ключевых проблем данного подхода. Эффективность разработки напрямую связана с наличием больших и сверхбольших ресурсов – размеченных корпусов текстов, онтологий и тезаурусов. Весьма важным является аспект стандартизации разработки, и в настоящее время де-факто сложился ряд стандартов, например стандарт WordNet[2] для лексических онтологий или стандарт PennTreeBank[3][4] для синтаксически размеченных корпусов текстов и др.

Другой проблемой является оценка эффективности используемых эмпирических критериев. Метрики числовых оценок в LE подобны хорошо известным в системах извлечения информации понятиям «точность» (precision) и «полнота» (recall). В основе получения оценок лежит сравнение результатов работы человека-аналитика и компьютерной программы при решении определённой задачи. Следует отметить, что области применения сравнительных оценок в LE постоянно расширяются.

Возрастающее использование статистических методов в задачах LE порождает некоторый отход от методов исследования и моделирования глубинных механизмов, лежащих в основе мышления и языка человека. Статистические методы в NLP позволяют достигнуть определённых результатов в решение ряда задач (распознавание речи, разрешение многозначности, аннотирование текстов и др.), однако представляются перспективным использование гибридных моделей, в которых используется различная техника, в том числе интроспективные методы.

Одним из перспективных направлений исследований в области извлечения информации (Information Extraction – IE) является направление «машинного обучения». Компьютерные системы, реализующие методы ML, ориентированы на получение новых знаний в результате автоматизации процесса обучения. Методы автоматического получения новых знаний на основе эмпирических данных можно успешно применять для формирования баз знаний. Это обстоятельство делается актуальными исследования в области обучения языку (Language Learning), результаты которых применимы в практических приложениях NLP-систем. Можно указать несколько причин, по которым исследования по ML становятся полезными в разработках NLP.

1. Сложность задач. Язык является сложноорганизованным объектом. Полная модель языка представляет сложное взаимодействие регулярностей, нерегулярностей, зон исключений и других явлений. Разработка такой модели может быть начата с разработки моделей отдельных подъязыков, описывающих относительно простые семантические области (например, медицинская диагностика и т.п.)
2. Реальные приложения. В настоящее время существует огромный рынок NLP – приложений (машинный перевод, реферирование и др.) Методы ML несомненно могут быть полезны в решение ряда важных проблем NLP-систем.
3. Доступность больших ресурсов данных. Стандартизация и открытость многих важных ресурсов обеспечивает необходимую ресурсную составляющую методов ML.

Распознавание именованных объектов (NER) назначает тег именованной сущности указанному слову, используя правила и эвристику. Именованный объект, представляющий человека, местоположение и организацию, должен быть распознан. Распознавание именованных объектов - это задача, которая извлекает номинальную и числовую информацию из документа и классифицирует слово на человека, организацию или категорию даты. NER классифицирует все слова в документе на существующие категории и «ни один из вышеперечисленных» [5].

Распознавание биомедицинских названных сущностей очень важно при языковой обработке биомедицинских текстов, особенно при извлечении из документов информации о белках и генах, таких как РНК или ДНК. Поиск названных объектов генов из текстов является очень важной и сложной задачей . Поиск имени гена в текстах соответствует поиску названия компании или имени человека в газетах. Распознавание биомедицинских именованных сущностей представляется более сложным, чем распознавание нормальных именованных сущностей. Многочисленные исследования позволили выявить названные объекты с помощью алгоритмов обучения под наблюдением, основанных на многих правилах [6].

Подходы к обучению с использованием контролируемых методов используют модели Маркова, деревья решений, метод опорных векторов (SVM) и условные случайный поля (CRF). Методы обучения с учителем обычно обучаются с использованием многих функций, основанных на различных лингвистических правилах, и оценивают эффективность с помощью тестовых данных[7].

Распознавание именованных объектов (NER) классифицирует все незарегистрированные слова, встречающиеся в текстах, и является подзадачей для извлечения информации. Обычно NER использует восемь категорий: местоположение, человек, организация, дата, время, процент, денежная стоимость и «ничего из вышеперечисленного». NER сначала находит именованные сущности в предложениях и объявляет категорию сущностей[8].

Распознавание именованных объектов имеет три подхода - на основе словаря, на основе правил и на основе машинного обучения. Подход на основе словаря хранит как можно больше именованных сущностей в списке, называемом справочником. Этот подход кажется очень простым, но в то же время имеет ограничения. NER сложен, потому что целевые слова в основном являются собственными существительными или незарегистрированными словами. Кроме того, новые слова могут генерироваться часто, и даже один и тот же поток слов может распознаваться как разнообразные именованные объекты с точки зрения их текущего контекста. Второй подход NER - подход, основанный на правилах [9]. Этот подход обычно зависит от правил и шаблонов именованных объектов, появляющихся в реальных предложениях. Хотя подходы, основанные на правилах, могут использовать контекст для решения проблемы нескольких именованных объектов, каждое правило должно быть написано до его фактического использования. Третий подход, основанный на машинном обучении, присваивает именованные объекты словам, даже если слова не перечислены в словаре, а контекст не описан в наборе правил. Для этих подходов в основном используются метод опорных векторов (SVM), скрытые Марковские модели, максимальные энтропийные Марковские модели и условные случайные поля (CRF) [9].

Исследователи по обработке естественного языка были заинтересованы в извлечении информации из генов, рака и белка из биомедицинской литературы [10]. Распознавание биомедицинских названных объектов, которое необходимо для извлечения биомедицинской информации, рассматривается как первый этап интеллектуального анализа текста в биомедицинских текстах. В течение многих лет признание технических терминов в области биомедицины было одной из самых сложных задач в обработке естественного языка, связанной с биомедицинскими исследованиями [11].

Биомедицинская NER сталкивается с трудностями по пяти причинам. Во-первых, из-за текущих исследований количество новых технических терминов быстро увеличивается. Очень сложно создать справочник, который включает все новые термины. Во-вторых, одни и те же слова или выражения могут быть классифицированы как объекты с разными именами с точки зрения их контекста. В-третьих, длина объекта довольно велика, и объект может включать контрольные символы, такие как дефисы (например, «12-*о*-тетрадеканоилфорбол 13-ацетат»). В-четвертых, выражения аббревиатуры часто используются в биомедицинской области, и они испытывают двусмысленность смысла. Например, «TCF» может относиться к «Т-клеточному фактору» или «Тканевая культуральная жидкость» [12]. Наконец, в биомедицинских терминах нормальные термины или функциональные термины объединяются, по-

этому биомедицинский термин может стать слишком длинным. Например, «HTLV-I-инфицированный» и «HTLV-I-трансформированный» включают нормальные термины «I», «инфицированный» и «трансформированный». Биомедицинскому NER трудно сегментировать предложение с именованными объектами. Изменения правописания также создают проблему. Кроме того, именованный объект одной категории может включать в себя другой именованный объект другой категории [13].

Методы машинного обучения, хотя и находят всё большее применение для различных задач обработки текстов, пока ещё остаются чрезвычайно сложными и трудоёмкими для реального применения. Это объясняется не столько сложностью алгоритмов обучения, сколько, возможно, неудачными методологическими подходами к обучению. Задачи обучения применяются фрагментарно, к какому-либо отдельному этапу последовательного процесса обработки текста. Именно поэтому приходится заниматься ручного разметкой, а не использовать результаты предыдущего обучения системы на предшествующих и взаимосвязанных этапах обработки.

Машинные методы обучения концептуальным знаниям представляют собой модель правдоподобных индуктивных и дедуктивных рассуждений, в которых вывод знаний н их использование не отделимы друг от друга. Реализация обучения в режиме правдоподобных рассуждений позволит организовать взаимодействие не только данных и знаний в процессах обработки текстов, но и моделировать процесс взаимодействий учителя и ученика в процессе приобретения знаний в схемах многоагентных взаимодействий.

ЛИТЕРАТУРА

1. Искусственный интеллект в медицине: главные тренды в мире // URL: https://medaboutme. ru/zdorove/publikacii/stati/sovety_vracha/iskusstvennyy_intellekt_v_meditsine_glavnye_trendy_v_ mire/ (Дата обращения: 30.09.2019).
2. Wordnet. // URL: http://wordnet.prin eton.edu/. (Дата обращения: 30.09.2019)
3. The Penn Treebank. // URL: http://www. is.upenn.edu/ ~ treebank/ (Дата обращения: 30.09.2019)
4. Mar us M., Santorini B., Mar inkiewi z M.A. Building a large annotated orpus of English: The Penn Treebank – Comput. Linguist. 1993. V. 19, No 2. P. 313 330.
5. Sang EFTK, Meulder FD. Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. In: Proceedings of the seventh conference on natural language learning at HLT-NAACL, vol. 4. 2003. p. 142–7.
6. Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th international conference on computational linguistics. Association for Computational Linguistics, vol. 1. 2002. p. 1–7.
7. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008;13:652–63.
8. Wilbur J, Smith L, Tanaben L. Biocreative 2 gene mention task. In: Proceedings of second BioCreative challenge evaluation workshop. 2007.
9. Rau LF. Extracting company names from text. In: Proceedings of the conference on artificial intelligence applications of IEEE, vol. 1. 1991. p. 29–32.
10. Zhao S. Named entity recognition in biomedical texts using an HMM model. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics. 2004. p. 84–7.
11. Sekine SN. Description of the Japanese NE system used for Met-2. In: Proceedings of the message understanding conference. 1998. p. 1314–9.
12. Lee KJ, Hwang YS, Rim HC. Two phase biomedical NE recognition based on SVMs. In: Proceedings of the ACL 2003 workshop on natural language processing in biomedicine. Association for Computational Linguistics, vol. 13. 2003. p. 33–40.

13. Song Y, Kim E, Lee GG, Yi B. POSBIOTM-NER in the shared task of BioNLP/NLPBA 2004. In: Proceedings of the international joint workshop on natural language processing in biomedicine and its applications. Association for Computational Linguistics. 2004. p.100–3.

ПРИМЕНЕНИЕ МЕТОДА КОНСТРУИРОВАНИЯ УПРАВЛЕНИЯ НА МНОГООБРАЗИИ К ЗАДАЧЕ ИММУНОЛОГИИ НА ПРИМЕРЕ МОДЕЛИ «ХИЩНИК-ЖЕРТВА»

С.И. Колесникова, М.Д. Поляк, В.А. Аврамёнок
(Санкт-Петербург, Санкт-Петербургский государственный университет аэрокосмического приборостроения)
skolesnikova@yandex.ru, m.polyak@guap.ru, _ava@outlook.com

APPLICATION OF THE CONTROL DESIGN METHOD ON MANIFOLDS TO THE IMMUNOLOGY PROBLEM ON THE EXAMPLE OF THE PREDATOR-VICTIM MODEL

S.I. Kolesnikova, M.D. Polyak, V.A. Avramyonok
(St. Petersburg, St.Petersburg State University of Aerospace Instrumentation)

Abstract. The problem of applying a control algorithm on a manifold to a stochastic nonlinear object represented as a system of nonlinear stochastic differential equations is considered. The applied interpretation of this description is a classical object from the field of knowledge immunology ("predator-prey"). The control variable characterizes the scheme of administration (daily) of donor antibodies / immunoglobulins into the body. A control synthesis algorithm based on the mathematical apparatus of the method of analytical design of aggregated controllers is proposed. The results of numerical modeling are presented, from which the operability of the obtained control system follows. The results obtained can be used for similar objects of higher dimension.

Keywords: Nonlinear stochastic object, robust regulator, target manifold, immunology problem, immune response control.

Введение в проблему Проблемы иммунологии всегда были актуальны, особенно в настоящее время, когда производительность вычислительных машин позволяет работать с достаточно сложными нелинейными моделями, трудоемкими с точки зрения вычислений.

Существует большое количество математических моделей (например, [1, 2]), описывающих процессы в иммунологии. Для иллюстрации применения принципов синергетической теории управления (СТУ) к исследованию объектов данной прикладной направленности будет рассмотрена модель типа «хищник-жертва», базовое описание которой представлено системой нелинейных дифференциальных [2-6] или разностных уравнений (полученных на основе дискретизации, в том числе).

Выбор модели обусловлен её фундаментальностью, поскольку модель «хищник-жертва» либо непосредственно применяется для описания объекта иммунологии, либо является составной частью большого числа математических моделей [5, 6], характеризуется хорошей изученностью и относительной простотой.

Суть СТУ [7] заключается в создании управляемой динамической декомпозиции нелинейных многомерных систем и направленной самоорганизации динамических системна основе искусственного конструирования притягивающих многообразий (разновидности множеств состояний с аттрактивным свойством) в фазовом пространстве в виде описания $\psi(x(t)) = 0, t \to \infty$, где $x = (x_1, ..., x_n)$ - вектор состояния объекта управления.

Цель данной работы - демонстрация указанной техники конструирования управления с интерпретацией в терминах иммунологии на простом примере объекта типа «хищник-жертва», являющимся составной частью многих систем (см. напр. [1-6]).